ISSN 1695-5498

HIPERTEXT.NET
Revista Académica sobre Documentación Digital y Comunicación Interactiva
DEPARTAMENTO DE COMUNICACIÓN · GRUPO DE INVESTIGACIÓN DIGIDOC

upf. Universitat Pompeu Fabra Barcelona

Número 20 · Mayo de 2020

Data Journalism: From Social Science Techniques to Data Science Skills

Adolfo Antón Bravo
Universidad Politécnica de Madrid
adolflow@infotics.es
https://orcid.org/0000-0003-2437-9863

Ana Serrano Tellería
Universidad de Castilla La Mancha
anaserranotelleria@gmail.com
https://orcid.org/0000-0003-1625-4411

Periodismo de datos: de las técnicas de las ciencias sociales a las habilidades de la ciencia de datos

ABSTRACT

In very few years in journalism, we have gone from looking at social science techniques, what was called precision journalism, to dealing with open data as a huge source of information that lead us to data journalism what connects with data science in the sense of using -again- scientific methods to extract knowledge and insights from structured data. This article offers an overview of that evolution and focuses on some prototypes that have emerged in this new journalistic ecosystem of data journalism, data visualization and data literacy.

KEYWORDS

Data Journalism; Data Visualization; Open Data; Data science;Precision journalism

RESUMEN

En muy poco años hemos pasado en el mundo del periodismo de mirar con atención cómo aplicar las técnicas de las ciencias sociales, lo que se denominó periodismo de precisión, a tratar con los datos abiertos como una basta fuente de información que nos lleva al periodismo de datos, lo cual conecta con la ciencia de datos en el sentido de usar, de nuevo, métodos científicos para extraer conocimiento y descubrir los entresijos de los datos estructurados. Este artículo ofrece una panorámica de la evolución del periodismo de datos y se centra en algunos prototipos que han emergido en este nuevo ecosistema de periodismo, visualización y alfabetización de datos.

PALABRAS CLAVE

Periodismo de datos; Visualización de datos; Datos abiertos; Ciencia de datos; Periodismo de precisión

Antón-Bravo, A. y Serrano-Tellería, A. (2020). Data Journalism: From Social Science Techniques to Data Science Skills. *Hipertext.net*, (20), 41-54. DOI:10.31009/hipertext.net.2020.i20.04

https://doi.org/10.31009/hipertext.net.2020.i20.04

1. Introduction

The practice of data journalism runs parallel to its experimentation. This natural tension converges in a long string of prototypes, depending on how much importance is given to one or the other field involved in the experimentation, depending on whether one or the other aspect is of more or less interest, or on whether one has more competence over others at the time of elaborating the journalistic piece.

We place the beginning of "modern" data journalism in 2009, sheltered by the open data portals, the popularity of *open source* and standardization of *HTML5*, the perfect scenario into which sources of data, tools and output formats arise. It is the time when Simon Rogers -now on *Google*- publishes his first piece in *The Guardian Data Blog*.

Shortly before, in December 2007, the "Eight Principles of Open Data" were published in Sebastopol, California, after a meeting organized by Tim O'Reilly -*O'Reilly Media*- and Carl Malamud, and sponsored by *Google*, *Yahoo* and *Sunlight Foundation* and signed by thirty people including Lawrence Lessig and Aaron Swartz -*Creative Commons*- or Adrian Hollovaty -*ChicagoCrime*, *EveryBlock*, *Django*-(Malamund, 2007). To the *Open Source Initiative* and its new impulse of standardization of open source licences contributed the popularity of free software and open technologies as the Web.

Philip Meyer was the founder of the **precision journalism** approach, the use of the methods of investigation from social sciences applied to journalism (Meyer, 2002). In 2009 the new brand **data journalism** aimed to take advantage of the data sources that grew up and the set of tools from computer science that allowed cleaning, curing and analysing data in order to create stories that will be published mainly in a web format. Evolution leads us to the journalist as a kind of data scientist, web designer, web developer, or computer specialist but what all of these knowledges mean is that different profiles arise and evolve to produce different approaches and formats in journalism.

In November 2012 Rogers was the editor of the section of *The Guardian Data.* He performed a TED where explains that "data journalism is the new Punk (...) you have 3 chords to play and form a band and now you have three datasets to create a story" (TEDx, 2012). The intervention began with the *The Clash*'s classic "London Calling". He also showed the second page of the first issue of the underground fanzine *Sideburns*, published in London in 1977, which in its cover led to *Stranglers* with the following sentence: "This is a chord. This is another. This is a third. Now form a band" (The first chord was LA; the second, MI; and the third, SOL.)

It is a good metaphor for what is usually answered when someone asks *what is data journalism?* and the answer is *data journalism is telling stories based on data*. Logically, stories are not *told* from data by itself but journalists analyze data in such a way that it is able to extract interesting –journalistic– stories to become journalistic products by following a specific pattern, the **method** of data journalism, such as the one proposed by Paul Bradshaw: 1) Compile; 2) Clean; 3) Context; 4) Combine; and 5) Communicate (Bradshaw, 2011).

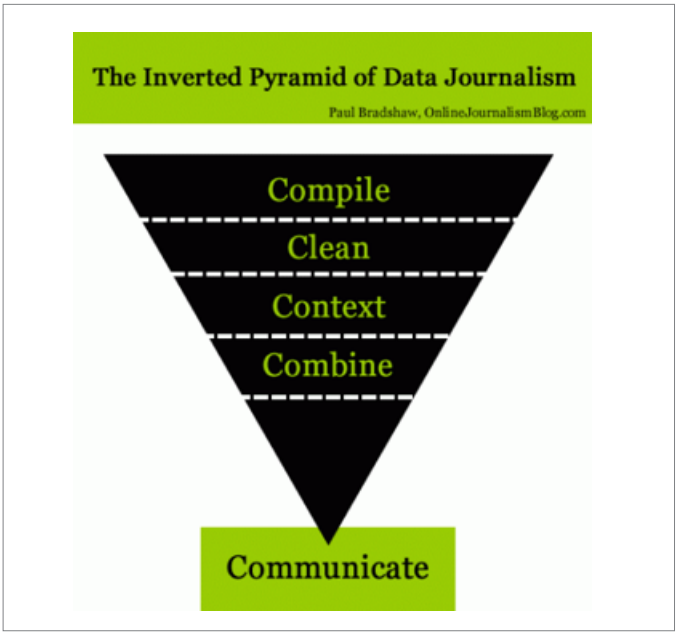

Figure 1. The Inverted Pyramid of Data Journalism. From: Paul Bradshaw, *Online Journalism Blog.com*.

The magic –associated with the ease– of the three notes to sing a song fits more in bands like *Sex Pistols* or *The Ramones* and probably with some easy reading simple stories. However *The Clash*'s punk is a higher level, something similar to the complexity of data journalism that comes from investigative reporting and the prototypes we introduce below.

Precisely this metaphor of the three chords is attributed to a country musician from Detroit called Harlan Howard, born in 1927 and well known for the album *To the silent majority with love* (Bonomo, 2015)[1], so titled two years after President Richard Nixon addressed the nation on 3 November 1969 in order to justify the Vietnam War, a speech known as *the silent majority* because of this paragraph:

> *Let historians not record that when America was the most powerful nation in the world we passed on the other side of the road and allowed the last hopes for peace and freedom of millions of people to be suffocated by the forces of totalitarianism. And so tonight-to you, the great silent majority of my fellow Americans-I ask for your support. I pledged in my campaign for the Presidency to end the war in a way that we could win the peace. I have initiated a plan of action which will enable me to keep that pledge. (Nixon, 1969).*

MC5 is one of the bands considered forerunners of punk or protopunk, founded in Detroit in 1965 two years before the 1967 Detroit rebellion where Philip Meyer tested the methods of social science research to apply them to journalism that he was studying at the *Niemen Lab* of the *University of Harvard*. If Philip Meyer released *Detroit Riots* in 1967, *MC5* released its album in 1968 *Kick out the Jams,* which included a version of *The Motor City is Burning* by John Lee Hooker (1967) who was talking about the city those days. Punk and data journalism from the roots (Anton-Bravo, 2013).

2. A Roundabout with the Terms

From the experience of the publication of *Detroit Riots* arose the precision journalism term to explain the methodology which lead the research. However it was not Meyer who invented the term. The draft of the book that Meyer was preparing had the title *The Application of Social and Behavioral Science Research Methods to the Practice of Journalism*, but Everette E. Dennis, then in the *Kansas State University*, was teaching a course at the *University of Oregon* about *The New Journalism* and mentioned Meyer's work as an *exotic* example of journalism "to contrast (his) scientific method with the artsy approach of those like Tom Wolfe and Jimmy Breslin who used short-story techniques to illuminate nonfiction" (Meyer, 1991). Before Meyer released his book with the *Indiana University Press Editorial* the term appeared in two others works: as a title of a chapter by Neil Felgenhauer of *The Magic Writing Machine* (Dennis, 1971) and in a footnote to *The New Journalism* (Johnson, 1971) referring to work of Ben Wattenburg, one of the co-authors of *The Real Majority*, an analysis of 1968 elections' electoral data.

Some people as Liliana Bonegru (s/d) places the beginning of *Computer Assisted Reporting, CAR,* before Meyer's work, in 1952, when the *CBS* tried to predict the results of the USA presidential elections using a computer. That was the first national broadcast -coast to coast- of a national election and CBS, with Walter Conkrite as master of the ceremonies, relied on a *Remington Rand UNIVAC* -Universal Automatic Computer- to predict the results between the Democratic candidate Adlai Stevenson and the Republican Dwight Eisenhower. All political analysts took Stevenson's victory for granted but with the 7% of the vote *UNIVAC* predicted an Eisenhower's landslide victory, which created a stir that even lasted until the definitive results were known. The success of the prediction did not do the computer more popular at journalists' eyes. Conkrite explained that they used it as a complement to their coverage, not as the main part of his investigation:

> *We saw it as an added feature to our coverage that could be very interesting in the future, and there was a great deal of pride that we had this exclusively. But I don't think that we felt the computer would become predominant in our coverage in any way.* (Goff, 1999)

For Bonegru it is the work with computers and data what determines to be CAR, and she may be right, but it wasn't until Meyer's *Detroit Riots* research that it was not introduced in a habitual way in the daily journalistic work. In the USA the term gives the name to the *National Institute of Computer-Assisted Reporting*, created in 1989 fourteen years after the *IRE*, *Investigative Reporters and Editors*, was founded to study investigative journalism. (Bonegru, s/d).

It wasn not until the 1990s' that precision journalism did not came to Spain. José Luis Dader and Pedro Gómez had attended to various events in the USA and give a talk at *El País* with the title *The development of 'precision journalism' in the USA* (Dader y Gómez, 1991). It is very interesting and revealing the antetitle, *A new socio-statistical information,* as they appealed directly to the use of social science techniques as Meyer and other had begun to do. How can we explain the lack of development of this kind of journalism in Spain when it seemed to be pretty interesting for investigative reporting? Looking back there could be three factors at least to explain it.

Firstly, the tone. Dader and Gómez begin their article pointing out that the census that the *INE -Instituto Nacional de Estadística*, *Spanish National Institute of Statistics*-, intended to carry out in 1991 came up against with some setbacks that qualify as "a wave of third worldism, under the banner of a misunderstood safeguard of the privacy" (which) "snatched the heads of many politicians and intellectuals, urging citizens not to reveal any part of the requested data" (Dader y Gómez, 1991). I don't know if they refer to the meaning of *critical Third World* or the one picked up by the *Royal Academy of Spanish Language*, condition of *third world*: "1. adj. belonging or relating to the third world; 2. adj.

depect. very high quality deficient". What is relevant is the different consideration of personal data in the USA and Spain -still a tremendously contemporary issue– and also of (lack of) access to information and (lack of) transparency laws.

Secondly, Dader and Gómez were not active journalists like Meyer and others who were looking for improvements to their profession. They came from the academia where the journalistic practice was far from having a connection as strong as the one that might be presupposed in the USA according to the readings of those texts and bibliography.

Thirdly, it makes sense to believe that they perform a biased reading of the term and the practice, turning it into something like what the character *Professor Keating* criticized in the fragment *Understanding poetry* of the *Dead Poets Society*, focusing on the *socio-informatics path*" as a *boring* object to study instead of as a practice sticked to investigative and CAR.

More academic language can be read in the article of Fermín Galindo Arranz *Propuesta de periodización histórica y evolución conceptual del periodismo de precisión* (2004), where establishes a difference between investigative journalism and precision journalism as the intention of this is:

> *(…) to create an objectified and systematized body of knowledge that surpass the conventional stereotype that the journalistic research is a matter of particular intuition or stroke of fortune in the reception of some revelations, or of the non-transferable "journalistic olfaction." With these three characteristics (initiative of the journalist, topic of general relevance and unveiling of a secret) investigative journalism is distinguished from the simple leaks* (Galindo-Arranz, 2004)

We are concerned that similar mistakes have been made today that can explain the slow unfolding and adaptation of the data journalism in Spanish newsrooms. In any case there are some projects that carry out the flag of the data journalism, some of them are below.

Regarding the term data journalism, Wikipedia distinguishes the data journalism from the *data driven journalism* and also from the *database journalism*. According to this definition, data journalism focuses on "the increased role that numerical data is used in the production and distribution of information in the digital era. It reflects the increased interaction between content producers (journalist) and several other fields such as design, computer science and statistics". That is to say, relates the term to the knowledge involved: journalism, design, programming and statistics.

This set of overlapping competencies coming from various sciences is also the subject of the five tips offered by Troy Thibodeaux in *Poynter* (2011), where he begins by naming some of the terms used to define the person who performs the task data journalism, which is itself another source of debate: *data journalist*, *computer-assisted reporter*, *newsroom developer* or *journo-geek* are some of them. Thibodeaux wonders if the terminological indeterminacy of the people who does data journalism has an impact on the understanding of the practice itself. Thibodeaux puts the accent on considering data journalism as a set of competencies that come from diverse and overlapping knowledge:

> *We have the statistical methods of social scientists, the mapping tools of GIS, the visualization arts of statistics and graphic design, and a host of skills that have their own job descriptions and promotion tracks among computer scientists: Web development, general-purpose programming, database administration, systems engineering, data mining (even, I hear, cryptography). And the ends of these efforts vary as widely as their means: from the more traditional text CAR story to the interactive graphic or app; from newsroom tools built for reporters to multi-faceted websites in which the reporting becomes the data.* (Thibodeaux, 2011)

What means, at least: 1) Statistical methods of the social sciences; 2) Management of geographic information system tools -GIS, maps-. ; 3) Ability to visualise statistics; 4) Ability to provide an attractive graphic design to the projects; 5) Web development; 6) Some programming skills; 7) Knowledge of database systems; 8) Notions of Data Mining; and 9) Practice with cryptography.

Skills	Provenance	Tools
Social science	Methods of the Social Sciences	Statistics, polls, interviews
Cartography	Geographic Information Systems	GIS, PosGIS, QGIS
Statistics	Ability to visualise statistics	R
UX	Ability to provide an atractive design to the projects	Bootstrap, CSS
Web	Web development	HTML5, CSS3, JavaScript

Skills	Provenance	Tools
Programming	Programming skills	Bash, Python, R, JS…
Databases	Knoledge of database systems	SQL, MySQL, PosgreSQL...
Data mining	Notions of Data Mining	KDD, Statistics, Maths
Cryptography	Practice with cryptography	GPG, PGP, Unix

Table 1. List of skills proposed by Thibodeaux and their correspondant tools

The output formats vary and conform, according to Thibodeaux, some **prototypes**:

- Classic CAR stories
- Interactive graphics
- Applications
- Tools for writing
- Websites

Another issue raised by Thibodeaux to explain the difficulties in defining data journalism comes from a lack of definition of what is considered *data:*

> *After all, anything countable can count as data. Anything that a computer processes is data. So, on some level, all journalism today is data journalism (certainly it's all "Computer Assisted"). Real data journalism comes down to a couple of predilections: a tendency to look for what is categorizable, quantifiable and comparable in any news topic and a conviction that technology, properly applied to these aspects, can tell us something about the story that is both worth knowing and unknowable in any other way.* (Thibodeaux, 2011)

If everything in the work with a computer can be considered *data* and everything a computer processes is *data*, then every journalist work driven by a computer (assisted by a computer) would be data journalism? It is not so simple as Thibodeaux considers that data journalism has two characteristics of its own:

1. Looking for patterns or curiosities among what is categorizable, quantifiable and comparable.
2. It takes as a premise that the use of technology can lead us to obtain an informative value that is not possible to find it in any other way.

And it concludes with a recommendation: do not care what you know about the domain -of data journalism- or do not mind how to write the story -with the new tools-, what journalists need to know is that they are storytellers and, therefore, either through the words or pixels, *write*!.

But we also have a clear goal: we're storytellers, through word or pixel, and the story won't wait for us to finish our self-imposed curriculum. So, pick up what's at hand, learn what you need to get to the next step in your project and get to something real as soon as possible.

In 2011 when Rogers summarized his first two years of Datablog's life he remained Adrian Holovaty's (2009) answer to the question he was asked about whether what he did was data journalism or not.

Is data journalism? Is it journalism to publish a raw database? Here, at last, is the definitive, two-part answer:

1. Who cares?
2. I hope my competitors waste their time arguing about this as long as possible.

Paul Bradshaw asked the same question and answered in that digital sense: all the information is reduced to numbers, to 0 and 1, bits. Therefore, "the data are not only the source but also the tool with which we tell a story, or it can be both" (Bradshaw, 2010).

In the set of fields or skills some people like Henk Ess do not include in the equation the process or the tools of design and visualization, he puts the focus of data journalism on the data processing that leads to a story[2]. Not having into account visualization as part of the analytic stage is one of the typical errors of many data journalism projects as far as data analysis it is also performed with visualization tools.

Regarding the open data, the pioneers of the *Guardian Data Blog* did something that can be named *Open Data Journalism*. The adjective *open* is very important because it does not points out only the fact that open data can be used freely as a data source but also because the journalistic product is open data in the sense of that the data sources and the methodology are published and shared with open licences. The *Datablog* hosted in *Google Spreadsheet -Google Drive-* all the

data sheets while *FiveThirtyEight*, *SRF Data* or *El Confidencial* use *Github*.

FiveThirtyEight publishes its data hosted in *Github* with the source code and licensed with *Creative Commons Attribution 4.0 International* for data and *MIT* licence for code. The data journalism unit of Swiss public television, *SRF Data*, has a repository in *Github* with the code and the data of 38 projects, tests or the web itself. Similarly, *El Confidencial*'s *Data Unit* shares some their data sources in *Github*.

El Confidencial's *Data Unit* is part of the Spanish data journalism community which was born in November 2011 when a group of people that come from journalism, data visualisation, access to information initiatives, transparency or open data community create the "Data journalism working group" at *Medialab* (Madrid, Spain) leaded by Mar Cabra in the first stage and Adolfo Antón Bravo until its end (May 2019). They organised several working sessions, seven editions of the data journalism production workshop, data journalism conferences and other activities.

In the presentation of the first *Open Data and Data Journalism Conference*, Karma Peiró, the leader of the activities in Barcelona (Catalonia), acknowledged that the main objective of the event was:

> *to awake the interest in data journalism, which mixes professional profiles (computer scientists, programmers, journalists, engineers…) designers, etc.) and put them to work together. The challenge is that we can learn from each other, that every one of us here, when we say goodbye, let's know a little more. That we work in a network to project new cases of data journalism. (…) The aim of the organization is to open a way that allows to increase the democratization of information, to make it much more accessible, to let it be out of the classic media corsets, to open new horizons so that we can all understand what's going on right now.* (Peiró, 2013)

Just in December 2013 it was published the *Law of Transparency, access to public information and good governance* (BOE, 2013), which began to be applied in December 2014 in institutions and in 2015 in the autonomous regions, a legal framework that brought us closer to Europe.

3. Prototypes

England, August 2011. A young is killed by police in Tottenham that unleashes a wave of protests and riots in the streets on the periphery of the main cities. Meyer's *Detroit Riot* was an inspiring work to check again so he was invited to comment the ongoing large-scale social science investigative research over the riots leaded by *The Guardian* and the *London School of Economics*, *Reading the Riots*. He pointed out that the work that the Guardian was publishing was on a much larger scale and that there was also "the first journalistic application of grounded theory that I know about. (…) as in Detroit, the Guardian's historic contribution is in the method, not the machinery" (Meyer, 2011).

Following the five templates proposed by Thibodeaux (classic CAR stories, interactive graphics, applications, tools for editorial and web sites that vary in appearance depending on the subject matter) we add a few more with specific examples. This list ends up with thirteen prototypes:

1. Applications or Newsapps (newsapps)
2. Interactive tools (interactives)
3. Classic stories CAR (Computer Assisted Reporting) or "big projects" (large)
4. Tools for writing or support for collaborative projects. (supporg)
5. Websites with a set of articles (set of articles)
6. Projects that rely heavily on data visualisation. (datavis)
7. General purpose applications (general purpose apps)
8. Public service tools (public service)
9. Projects where maps are a central element. (maps)
10. Single-page projects with horizontal or vertical scrolling. (one-page project)
11. Projects that support or use video as a central element. (video)
12. Leaks (leaks)
13. Social sciences approach. (precission journalism)

N.	Thibodeaux (5)	Antón and Serrano (13)
1	CAR	CAR
2		Leaks
3		Social Science Approach
4	Interactive Graphics	Interactive tools
5		Datavis
6		Maps
7	Apps	Apps or Newsapps
8		General Purpose Apps
9		Public Service Tools
10	Tools for Writing	Tools for writing or support for Collab. projects

N.	Thibodeaux (5)	Antón and Serrano (13)
11	Websites	Websites with a set of articles (documents)
12		Single Pages (One page)
13		Multimedia based website

Table 2. Prototypes: Thibodeaux vs Antón and Serrano

3.1. Newsapps

ProPublica proposes *Newsapps* which is what the applications of news, graphics, databases and tools they create with respect to a particular topic.

Nonprofit explorer

This ProPublica project[3] has a data exploration tool to search over 3 million records of entity tax exemptions and details such as economic compensation, benefits and expenses. Being such a volume of data, they have created an API and offer an embedded search engine to be embed in any other website.

Nursing Homes

With *Nursing Homes*[4] ProPublica compares nursing homes in each state based on the deficiencies observed by the regulators and the fines imposed in the last 3 years. You can search over 60,000 reports of inspection. They have created a guide to search the data, stories like *Two Deaths, Wildly Different Penalties: The Big Disparities in Nursing Home Oversight* and even a zip file that opens the data they got.

Elecciones 2018

One of the pioneering teams in Spanish, *La Nación Data*, created *Elections 2018*[5] to concentrate the special contents of La Nación's Data Intelligence Unit related to the 2018 electoral campaign.

3.2. Interactives

There are several possibilities of interaction in the Web environment but probably FiveThirtyEight makes its best in this kind of interactive news that highlights some data from the whole picture as well as a careful aesthetics.

Club Soccer Predictions

Club Soccer Predictions[6] forecasts *Soccer Power Index* (SPI from ESPN) ratings for 36 leagues, updated after each match. They have updated and adapted SPI to incorporate soccer club's data going back to 1888 (from more than 550,000 matches in all) that we've collected from *ESPN's* database and other soccer data sources as well as from play-by-play data produced by *Opta* that has been available since 2010.

Congress Generic Ballot Polls

In this other project from *FiveThirtyEight* it is shown up the dates of tracking polls from the same pollster overlap, only the most recent version is shown. *Congress Generic Ballot Polls*[7] list those that ask either which party's candidate a respondent would vote for in his or her district or which party the respondent would prefer control Congress.

Mapas del descontento

The Spanish artist Martín Nadal developed *Mapas del descontento*[8] (Map of Social Protest) along the workshop *Visualizar Open Cities* (2015) and eventually at *EditorsLab* (2015). It is a data-driven journalism project that aims at showing a broad panoramic of the social protest in Spain from 1976 to 2015. The data was gathered from the tag explorer of the Spanish newspaper *El País*. To get the information, the tag *Social protests* was selected and data was stored through an automatic scraping process getting more than 2,500 articles where titles, keywords, places were analyzed.

3.3. Large Projects

Some of these projects were launched from the *ICIJ*, an international network independent of journalists, based in Washington, founded in 1997 by the *Center for Public Integrity* but not related to it since 2017, when it became independent.

Panama Papers

The Panama Papers[9] was a huge collaboration project of more than 100 media partners, including members of *Organized Crime and Corruption Reporting Project* (*OCCRP*). Thanks to this project, the following were launched research in more than 82 countries. In Spain, the main consequence was the resignation of the Minister of Industry, José Manuel Soria the year Spain leaded the open data hype with the celebration in Madrid of the *Open Data Conference IODC16*. Other projects that conjugate filtrations and capital flight: *SwissLeaks*, *LuxLeaks* or *Offshore Leaks*.

Dollars for Docs

Dollars for Docs[10] is a long-distance *ProPublica* project started in 2013 with updated data in 2016. Sample payments made by medical and pharmaceutical consulting firms to doctors. Given the size of the project and the

set of articles published, they explain not only how they got some data but also their analysis methodology and offer data to be downloaded.

Migrant Files

The Migrant Files[11] was a project supported in part by *Journalismfund.eu* to show up the number –and eventually the identity– of thousand of migrants trying to reach Europe crossing the Mediterranean Sea. It was discontinued on June 24, 2016 when they updated the database for the last time after reaching the goal they set.

3.4. Support for Collaborative Projects

Following the principle of collaboration of data journalism community there have been a lot of tools and projects to engage newsrooms' teams or teams of journalists. Some of them are:

Document Cloud

One of the pioneering and reference tools for data journalism is *Document Cloud*[12], a platform founded on the belief that if journalists were more open about their sourcing, the public would be more inclined to trust their reporting. The platform is a tool to help journalists share, analyse, annotate and publish the source of documents.

CrowdNewsroom

CrowdNewsroom[13] is a platform to create and manage collaborative investigative projects with the help of the community. It is a platform founded by *Correctiv* to create and conduct collaborative investigations with the help of communities. It allows find sources and verify data. Through the platform, citizens share data and their personal story.

Datakit

The *Datakit*[14] is an open-source command-line tool by *Associated Press* designed to better structure and manage projects. It makes it easier to standardize and share work among members of your team, and to keep your past projects organized and easily accessible for future reference.

AP DataKit works off a basic framework that includes the core product and a few key plugins to help you manage where your data files and code are stored and updated.

3.5. Set of Articles

From big to small newsrooms the set of articles is a prototype widely used in the investigative journalism.

The Soccer Files

Reuters Investigates finds a pattern in "The Soccer Files"[15]: European football clubs whose owners are multimillionaires of Middle East like *Manchester City* and *Paris Saint Germain* have harvested millions of euros through sponsorship agreements with the citation good of the authorities in spite of the refusal of experts independent.

The Missing

The Missing[16] is a project from *Associated Press* to explore the stories of migration from across the world.

Mar Menor, historia profunda de un desastre

Mar Menor, historia profunda de un desastre[17] (Mar Menor, history of an environmental disaster) is an independent investigation of *Datadista* leaded by Antonio Delgado and Ana Tudela. They find out how it is possible to pollute uninterruptedly the lagoon along three decades of breaking the law, overexploiting the water that feed the irrigation system and a chaos of canals, desalination plants and streams that pour nitrates out of control.

3.6. Data Visualization

Alberto Cairo commented in an interview in 2012 (Garrido, 2012) that the data visualisation was a part of the data journalism process. That is to say, a visualization aspect to tell the story. However, since the data analysis tools from data science have been introduced in the practice of data journalism, visualisations are also taken into consideration in the analytical phase.

Festivals dominated by men

The *BBC England* data unit has performed projects such as "Festivals dominated by mal acts, study shows, as Glastonbury begins"[18] where they evidence the gender bias at festivals of UK music.

Soy de temporada (*seasonal fruit and vegetables*)

Soy de temporada[19] is a project developed within the framework of the workshop *Visualizar* "Migraciones" (2017) where migrations were held in so many ways.

How has my country voted at the UN

In *How has my country voted at the UN*[20] *Aljazeera* shows up how every country has voted at the *UN General Assembly* from 1946 to 2018.

3.7. General Purpose Applications

They can be carried out in an editorial office for internal use, for use or even carried out by people who do not belong to a newsroom.

LibreBOR

LibreBOR[21] -formerly known as *LibreBORME*- is a web platform for the consultation and analysis of the *Official Gazette of the Commercial Registry* (*Boletín Oficial del Registro Mercantil de España*), a newsletter published in PDF since 2009 and resilient to the Open Data movement. This tool adds the latest changes that are published and allows to make semantic searches, receive notifications and use these data.

Datashare

With all the experience from the data driven investigations at *ICIJ* they have developed *Datashare*[22] to analyze documents and extract entities as locations, companies or people. It is multiplatform as it works in a *docker*, it is free software and source code is available and very well documented.

ER Inspector

ER Inspector[23] is a way to find and evaluate every *Emergency Room*, it allows to look up hospitals ahead so everyone can evaluate where to go in an emergency. They get data on hospital quality measures, such as wait times, patient ratings and citations for emergency room violations.

3.8. Public Service

Projects that come from civil society entities, normally with a vocation of public service and that offer some service that does not perform the administration.

Sold From Under You

Sold From Under You[24] is a large-scale data-led collaborative investigation –*The Bureau of Investigative Journalism* and *HuffPost UK*– into the sell-off of public spaces by local authorities which revealed, for the first time, the scale to which the local government funding crisis is affecting public services, public spaces, and public servants.

Health Inspection in Madrid

The map of health inspections[25] in Madrid has become one of the most interesting stories in 2019. For the population of UK, New York or France this is available in the open data portal itself but in Madrid there has to be a Civil Society Organization which makes the work of enrich the data published in the open data portal.

Govern Obert *(Open Government)*

Govern Obert[26] aimed to explore the Catalan government through visualizations of Departments of the regional Catalan government -Generalitat- through an XML from it open government portal. The hierarchy between the different organization within each department was added by scraping. The code and data (csv or json, produced in Python) was published in Github.

3.9. Maps

Lot of projects where maps have a central or determining role. A map situate the story, get to know the story through the knowledge of places where other stories or history live.

España en llamas *(Spain in flames)*

España en llamas[27] was born as a research on forest fires from 2001 to 2010. The 170,822 fires burned 1,137,566 hectares (ha), as much as the entire Region of Murcia (South East). The map shows up the which ended up with 100 ha or more, 61.5% of the total burned area.

Tell-all telephone

Tell-all telephone[28] was one of the first and outstanding examples of journalism of data in Germany. The story was about *Green Party* politician Malte Spitz sued to have German telecoms giant *Deutsche Telekom* hand over six months of his phone data that he then made available to *ZEIT ONLINE*. They combined this geolocation data with information relating to his life as a

politician, such as Twitter feeds, blog entries and websites, all of which was *freely* available on the internet.

Vidas Contadas

During the first *Spanish Data Conference* (2013) a hackathon of projects was carried out and *Vidas Contadas*[29] (Counted lifes) was one of them. In the context of the historical memory of the executed, disappeared and reprisals from the fascist coup d'état, the three-year war and the 40-year national-catholic dictatorship, this project tried to map and count every victim. There are still today 100,000 people uncounted in Spain buried in mass graves in the sides of the roads. During *Visualizar* "Commons" (2015) it got another impulse.

3.10. One Page Project

A solution widely used in stand-alone projects, whether they are part of or not of an essay.

NSA Files: Decoded

The Guardian has published selection of classified NSA documents[30], passed on by whistleblower Edward Snowden. Some have been redacted to preserve author anonymity.

Dime cuánto cobras y te diré dónde vivir

Hosted under the domain alquilarenelcentrol.lol[31] *Dime cuánto cobras y te diré dónde vivir* (Tell me your wage and I'll tell you where you can afford to live) is a research conducted during the *Data Journalism Workshop La España vacía* (the empty Spain, 2017) which warns about the increase in the price of rents in Madrid and Barcelona at a much faster rate than salaries, so that it is finding housing has become an arduous challenge. Gentrification, tourist flats and speculation are also involved.

Historia de Zainab *(Zainab's tale)*

"Historia de Zainab"[32] was developed during the workshop *Visualizar Migraciones* (Migrations, 2017). It is like a tale where a fictitious Syrian girl named Zainab deals with challenges in her periplos to flee from war and travel to Europe.

3.11. Video Support

Video is the main element of this class although it is also combined with audio. The interest for audio is growing hence there will probably be an addition to this categorisation or a merge of this class. So far:

Pundering Cambodia's Forests

Pundering Cambodia's Forests[33] is a project developed by *AlJazeera* that puts the focus on video and maps to show changes produced in Cambodia's forests.

Población dirigida

Between 1939 and 1973, the National Institute of Colonization (INC) promoted the construction in Spain of more than 300 villages. The ambitious plan that aimed to create large irrigated areas and increase its productivity mobilized approximately 55,000 families. *Población dirigida*[34] (Directed Population) was developed along the workshop *Visualizar Migraciones* (Migrations, 2017).

This fact was the most important migratory movement promoted by the Spanish State in the twentieth century. Colonization was a multidimensional process characterized by a big data collection. Accessing this data is accessing the memory of a transformation. This is the story of a set of worlds created from nothing, narrated from the consultation and continuous visualization of historical data of archives and legitimate studies.

Working With Dark Light

Working With Dark Light,[35] by Washington Post, shows how Puerto Rican artists provide relief from hurricane Maria. How was life without power after hurricane Maria, the life in the dark.

3.12. Leaks

Filtrations are fundamental to (investigative) journalism and classic CAR, from the famous deep throat of the journalists from the *Washington Post* at the Watergate to the revelations of Julian Assange's *Wikileaks*. In the case of the *ICIJ*, these leaks have allowed the great world-wide investigations carried out and both they as well as other media have formalized the way to get anonymously leaks.

ICIJ

The *ICIJ* has a platform called "Leak"[36] to securely submit all forms of content that might be of public concern, information that relates to potential wrongdoing by

corporate, government or public service entities in any country, anywhere in the world. That's part of its job.

Filtrala

Filtrala[37] (leak it) is an independent platform for sending information of public interest to the media and civil society organizations in a safe and secure manner, anonymous. It is participated by *eldiario.es*, *Revista Mongolia*, *Civio*, *Porcausa*, *Ecologistas en Acción*, *Greenpeace*, *Facua* and the *Platform for Action. Freedom of Information (PDLI)*.

Wikileaks

Probably the first of the great leaks in the digital era. *WikiLeaks*[38] was founded by Julian Assange in 2006 when releases large datasets of censored or otherwise restricted official materials involving war, spying and opinions from the cables of U.S. Diplomacy.

3.13. Social Science Driven Projects. Precision Journalism

In this prototype there is only one project to mention: "Reading the Riots"

Reading the Riots

Reading the Riots[39] is a data-driven study developed by *The Guardian* and the *London School of Economics* to study the causes and consequences of the riots of 2011 in England. They have verified every incident; painted a map; run several community conversations; analyzed the cases; investigated data and eventually the 2.57m tweets sent around the riots themselves.

Furthermore, as a social research project they counted with a specially-recruited team interviewed around 270 people about the riots and why they had been involved. The project was the first time such a major attempt had been made to forensically examine the motivations behind a riot since the work in Detroit in 1967.

Proposal of Prototypes

Section	Project	Organisation	Country
Newsapps	Nonprofit explorer	ProPublica	USA
Newsapps	Nursing Homes	ProPublica	USA
Newsapps	Elecciones 2018	La Nación	Costa Rica
Interactives	Club Soccer Predicitions	Five Thirty Eight	USA
Interactives	Congress Generic Ballot Polls	Five Thirty Eight	USA
Interactives	Mapas del descon-tento	Medialab-Prado	Spain
Large	Panama Papers	ICIJ	USA/Inter-national
Large	Dollars for Docs	ProPublica	USA
Large	Migrant Files	Journalism-fund.eu	Internatio-nal
Support	Document Cloud	Document Cloud	USA
Support	Crowd-Newsroom	Correctiv	Germany
Support	Datakit	AP	USA
Set of Articles	The Soccer Files	Reuters	UK
Set of Articles	The Missing	AP	USA
Set of Articles	Mar Menor, historia de un desastre	Datadista	Spain
Datavis	Festivals dominated by men	BBC	UK
Datavis	Soy de tem-porada	Medialab-Prado	Spain
Datavis	How has my country voted at the UN	AlJazeera	Qatar
General Purpose Apps	LibreBOR	Individual	Spain
General Purpose Apps	Datashare	ICIJ	USA

Section	Project	Organisation	Country
General Purpose Apps	ER Inspector	ProPublica	USA
Public Service	Sold From Under You	The Bureau of Investigative Journalism and HuffPost UK	UK
Public Service	Health Inspection in Madrid	Civio	Spain
Public Service	Govern Abert	Individuals	Spain
Maps	España en llamas	Civio	Spain
Maps	Tell-all telephone	Zeit Online	Germany
Maps	Vidas contadas	Medialab-Prado	Spain
One-Page Project	NSA Files: Decoded	The Guardian	UK
One-Page Project	Dime cuánto cobras…	Medialab-Prado	Spain
One-Page Project	Historia de Zainab	Medialab-Prado	Spain
Video	Pundering Cambodias Forest	AlJazeera	Qatar
Video	Población dirigida	Medialab-Prado	Spain
Video	Working With Dark Light	Washington Post	USA
Leaks	Leak	ICIJ	USA
Leaks	Fíltrala	Associated WhistleblowingPress	Belgium
Leaks	Wikileaks	Wikileaks	International
Precission Journalism	Reading the Riots	The Guardian	UK

Table 3: List of projects classifed by section

4. Conclusions

Addressing an investigative data journalism project leads to clarify the object of study that we are dealing with, the data journalism, and requires a purposeful clarification to be made for the determine the scope of the proposal. Even more if it deals with the identification of prototypes –templates– that can be used in the better understanding of the projects being addressed and the idiosyncrasy and complexity of themselves in an evolutionary process of something which its definition it is based on scientific armor or more precisely a closed connection with data science.

To be honest we have passed over the diversity of definitions that the journalist who works on these projects can host. Undoubtedly the best definition of the above will contribute also to clarify the definition of the exercising of this work.

We have also assumed that the generic terms *data science* and *data scientist* are understood without any problem as long as the scientific method and the computer scientist tools are concerned in the work with data which includes artificial intelligence, machine learning, natural language processing, data mining, deep learning, knowledge graph, web scraping, web development among others.

Therefore, we get to the list of prototypes, thirteen, with at least three examples in each of them except social research topic. It is true that some of them could be in more than one category but that would mean another approach in this research. On the contrary, these thirteen prototypes reflect the broad spectrum of projects in the ecosystem of the data journalism. In any case it is a work in progress that the data journalism community will approve or not regarding not only the projects to be released or the reviewing of the projects already published but also with the reception and/or use of new tools or technologies that may emerge in the short term.

End Notes

1. There was also a song called "Silent Majority" against the war by Eddie Harris & Gene McDaniels (Live at Newport), https://www.youtube.com/watch?v=yjVICI49KHw
2. It is interesting to see "History of Data Journalism" that begins with Philip Meyer in 1970, when precision journalism is defined. https://www.youtube.com/watch?v=ItPTs48qcek
3. ProPublica, API Non profits, https://projects.propublica.org/nonprofits/api
4. ProPublica, Nursing homes. Nursing homes' guide https://www.propublica.org/article/whats-new-in-nursing-home-inspect; Nursing Home Oversight, https://www.propublica.org/article/two-deaths-different-penalties-disparities-in-nursing-homes-oversight; and Nursing home data, http://downloads.cms.gov/files/Full-Statement-of-Deficiencies-May-2019.zip

5. La Nación. Elecciones 2018. https://www.nacion.com/gnfactory/investigacion/2018/elecciones_presidenciales/diputados/portada.html
6. FiveThirty Eight, Club Soccer Predictions' methodology. ABCNews. https://fivethirtyeight.com/methodology/how-our-club-soccer-predictions-work
7. FiveThirtyEight. Congress Generic Ballot Polls. ABC News. https://projects.fivethirtyeight.com/congress-generic-ballot-polls
8. Martín Nadal, Maps of Discontent. http://mapas.muimota.net Se also, Martín Nadal, data artist, http://martinnadal.eu
9. Panama Papers. https://www.icij.org/investigations/panama-papers
10. Dolars for Docs. https://projects.propublica.org/docdollars. See also the code of Dolars for Docs: https://www.propublica.org/nerds/the-coders-cause-in-dollars-for-docs
11. The Migrant Files. http://www.themigrantsfiles.com
12. Document Clound. https://www.documentcloud.org
13. CrowdNewsroom. https://correctiv.org/crowdnewsroom-pro
14. AP DataKit. http://datakit.ap.org
15. The Soccer Files, A Reuters Series. https://www.reuters.com/investigates/section/soccer-files
16. AP News. The Missing. https://apnews.com/TheMissing
17. Mar Menor, historia profunda de un desastre. Datavista. https://datadista.com/medioambiente/desastre-mar-menor
18. BBC England Data Unit. https://github.com/bbc-data-unit. And Festivals dominated by male acts, study shows, as Glastonbury begins, by Sherlock & Bradshaw. https://www.bbc.com/news/uk-england-40273193
19. Soy de temporada. https://soydetemporada.es
20. How has my country voted at the UN. Aljazeera. https://interactive.aljazeera.com/aje/2019/how-has-my-country-voted-at-unga/index.html
21. LibreBOR. https://librebor.me
22. Datashare. International Consortium of Investigative Journalism. https://datashare.icij.org
23. ER Inspector. ProPublica. https://projects.propublica.org/emergency
24. Sold From Under You. The Bureau of Investigative Journalism. https://www.thebureauinvestigates.com/stories/2019-03-04/sold-from-under-you
25. Health Inspection in Madrid. Civio. https://civio.es/tu-derecho-a-saber/2019/10/10/consulta-los-locales-de-tu-barrio-que-suspendieron-en-la-ultima-inspeccion-sanitaria
26. Govern Obert. Generalitat de Catalunya. https://opengov.cat
27. España en llamas. Civio. https://civio.es/espana-en-llamas
28. Tell-all telephone. Zeit Online. https://www.zeit.de/datenschutz/malte-spitz-data-retention
29. Vidas Contadas. http://vidascontadas.org
30. NSA Files: Decoded. The Guardian.com https://www.theguardian.com/world/interactive/2013/nov/01/snowden-nsa-files-surveillance-revelations-decoded#section/1
31. Dime cuánto cobras y te diré dónde vivir. Medialab-Prado. https://alquilarenelcentro.lol
32. Historia de Zainab. PorCausa and Medialab-Prado. http://historiadezainab.org
33. Pundering Cambodia's Forests. Aljazeera. https://interactive.aljazeera.com/aje/2019/plundering-cambodias-forests/index.html
34. Población dirigida. Lxs colonxs de la España verde de Franco. https://territoriodedatos.org/poblacion-dirigida
35. Working with Dark Light. The Washington Post. https://www.washingtonpost.com/graphics/2018/national/puerto-rican-art-hurricane-maria/
36. Leak to us. ICIJ, International Consortium of Investigative Journalists. https://www.icij.org/leak
37. Filtrala. https://filtrala.org
38. Wikileaks. https://wikileaks.org
39. Reading the Riots. The Guardian. https://www.theguardian.com/uk/series/reading-the-riots

References

Antón Bravo, A. (2013). El periodismo de datos y la Web Semántica. *CIC, Cuadernos de Información y Comunicación, 18*(), 99-116. http://dx.doi.org/10.5209/rev_CIYC.2013.v18.41718

BOE (10 de diciembre de 2013). Ley 19/2013, de 9 de diciembre, de transparencia, acceso a la información pública y buen gobierno. *Boletín Oficial del Estado*, núm. 295. https://boe.es/buscar/act.php?id=BOE-A-2013-12887

Bonomo, J. (August 20, 2015). *Harlan Howard's Nixon's America. No such Thing as was*. http://www.nosuchthingaswas.com/2015/08/harlan-howards-crushing-on-silent.html

Bounegru, L. (s/d). Data Journalism in Perspective. *Datajournalism.com* https://datajournalism.com/read/handbook/one/introduction/data-journalism-in-perspective

Bradshaw, P. (2010). How to be a data journalist. *The Guardian.com*. https://www.theguardian.com/news/datablog/2010/oct/01/data-journalism-how-to-guide

Bradshaw, P. (2011). The inverted pyramid of data journalism. *OnlineJournalismBlog*. https://onlinejournalismblog.com/2011/07/07/the-inverted-pyramid-of-data-journalism

Dader, J.L. y Gómez, P. (December 03, 1991) El desarrollo del 'periodismo de precisión' en Estados Unidos. *El País.com*. https://elpais.com/diario/1991/12/03/sociedad/691714811_850215.html

Dennis, E. E. (1971). *The magic writing machine; student probes of the new journalism*. School of Journalism, University of Oregon.

Galindo Arranz, F. (2004). Propuesta de periodización histórica y evolución conceptual del periodismo de precisión. *Estudios sobre el mensaje periodístico, 10*(), 97-112.

Garrido, H. (May 16, 2012). Alberto Cairo: "La falta de interés en profundizar en el periodismo de datos no tiene que ver con los recursos ni con la crisis". *Open Datos Blog*. https://opendatos.blogspot.com/2012/05/alberto-cairo-la-falta-de-interes-en.html

Goff, L. (April 30, 1999). Univac predicts winner of 1952 election First for television and information technology. *CNN.com*. http://edition.cnn.com/TECH/computing/9904/30/1952.idg/index.html

Holovarty, A. (2009). *The definitive, two-part answer to "is data journalism?"* [Personal Blog]. http://www.holovaty.com/writing/data-is-journalism

Johnson, M. L. (1971). *The new journalism: the underground press, the artists of nonfiction, and changes in the established media*. University Press of Kansas.

Malamud, C. (2007). Open Government Working Group. *Public.Resource.Org*. https://public.resource.org/open_government_meeting.html

Meyer, P. (1991). *The New Precision Journalism*. https://carolinadatadesk.github.io/pmeyer/book

Meyer, P. (2002). *Precision journalism: a reporter's introduction to social science methods*. Rowman & Littlefield Publishers.

Meyer, P. (December 9, 2011). Riot theory is relative. *The Guardian.com*. https://www.theguardian.com/commentisfree/2011/dec/09/riot-theory-relative-detroit-england

Nixon, R. (November 11, 1969). Nixon's 'Silent Majority' Speech. *Watergate.info*. https://watergate.info/1969/11/03/nixons-silent-majori-

ty-speech.html

Peyró, Karma(2013). Karma Peiró about data journalism [video interview]. *Centre de Cultura Contemporània de Barcelona, CCCBLAB*. http://lab.cccb.org/en/karma-peiro-about-data-journalism

Rogers, S. (July 28, 2011). Data journalism at the Guardian: what is it and how do we do it? *The Guardian.com*. https://www.theguardian.com/news/datablog/2011/jul/28/data-journalism

TEDx (2012). *Data-journalists are the new punks: Simon Rogers at TEDx Pantheon Sorbonne* [Video]. https://www.youtube.com/watch?v=h2zbvmXskSE

Thibodeaux, T. (2011). 5 tips for getting started in data journalism. *Poynter* [original web discontinued]. https://web.archive.org/web/*/http://www.poynter.org/how-tos/digital-strategies/147734/5-tips-for-getting-started-in-data-journalism

CV

Adolfo Antón Bravo. Universidad Politécnica de Madrid (UPM), holds a PhD in Information Sciences Studies and a Master's Degree in Pedagogical Aptitude at the Universidad Complutense de Madrid (UCM). He is currently a researcher in Ontology Engineering at the Ontology Engineering Group which belongs to the Department of Artificial Intelligence and coordinator of the Master's Degree in Journalism and Data Visualisation at the Universidad de Alcalá de Henares (UAH). He was in charge of the Datalab at Medialab-Prado its three years of life and before he worked as organizer, coordinator and/or curator of data-related activities since 2013 including Data Journalism and Open Data Spanish Conference, Data Journalism Production Workshop and Data Visualisation Workshop Visualizar at Medialab-Prado (Madrid City Council cultural centre of the arts). He has also been a master's professor at Universidad Carlos III-Agencia EFE 's Master in Agency Journalism (2018-2019); master's professor at Centro Universitario Villanueva's Master in Data Journalism at the (attached to the UCM); master's professor at Universidad Internacional de La Rioja's Master in Investigative Journalism, Data and Visualization (2015-2016); master's professor at Universitat Oberta de Catalunya's Master of Free Software (2010-2013) and data journalism advisor for the International Open Data Conference 2016 (Red.es, Ministry of Economy and Business of Spain). https://infotics.es; https://github.com/flowsta

Ana Serrano Tellería. Faculty of Communication. University of Castilla La Mancha. Associated Professor (ANECA), belongs as a postdoctoral researcher to the following centers: ASCA & Media Studies Department, Faculty Of Humanities, University of Amsterdam; OsloMet Digital Journalism, Oslo Metropolitan University; DIGIDOC, Pompeu Fabra University; LabCom.IFP, University of Beira Interior, Portugal; MESO, San Andrés University, Argentina, and Northwestern University, USA; Innovation in Digital Media, Carlos III University and Innovamedianet: a network of researchers. Since receiving the "Extraordinary Ph.D Award" in 2012 for his thesis (2010) "Initial Node Design on Cybermedia: A Comparative Study", her academic and professional work has focused on cross / multi / transmedia communication; interface design; media literacy; media ecology; media studies; mobile devices; privacy; performative and visual arts. With more than 90 publications, participation in projects and member of national and international research and innovation laboratories; she has worked as a journalist in the media and communication offices; media consultant; head of research projects; actress, singer and assistant director; artist manager; manager of cultural projects and cooperation as well as event organizer. She has obtained relevant national and interntional scholarships. https://es.linkedin.com/in/ana-serrano-tellería-3b762322

OCM Observatorio de Cibermedios

https://observatoriocibermedios.upf.edu/

upf. Universitat Pompeu Fabra Barcelona

Departamento de Comunicación Grupo DigiDoc

El **Observatorio de Cibermedios** es una producción del *Grupo de Investigación en Documentación Digital y Comunicación Interactiva* (DigiDoc) del **Departamento de Comunicación** de la **Universitat Pompeu Fabra**.

El Observatorio de Cibermedios (OCM) forma parte del proyecto del Plan Nacional "***Creación y contenido interactivo en la comunicación de información audiovisual: audiencias, diseño, sistemas y formatos***". CSO2015-64955-C4-2-R (MINECO/FEDER), Ministerio de Economía y Competitividad (España).